Coping with unbalanced class data sets in oral absorption models

Article (Accepted Version)

Coping with unbalanced class datasets in oral absorption models

Danielle Newby[a], **Alex.A. Freitas**[b] **and**, **Taravat Ghafourian**[a*]

[a]*Medway School of Pharmacy, Universities of Kent and Greenwich, Chatham, Kent, ME4 4TB, UK*

[b]*School of Computing, University of Kent, Canterbury, Kent, CT2 7NZ, UK*

*** Corresponding Author,** Email: T.ghafourian@kent.ac.uk; Tel +44(0)1634 202952; Fax +44 (0)1634 883927

Abstract

Class imbalance occurs frequently in drug discovery datasets. In oral absorption datasets, in the literature, there are considerably more of highly-absorbed compounds compared with poorly-absorbed compounds. This produces models that are biased towards highly-absorbed compounds which lack generalization to industry settings where more early stage drug candidates are poorly-absorbed. This paper presents two strategies to cope with unbalanced class datasets: Under-sampling the majority high absorption class and misclassification costs using classification decision trees. The published dataset by Hou et al (2007), which contained percentage human intestinal absorption of 645 drug and drug-like compounds, was used for the development and validation of classification trees using C&RT analysis. The results indicate that under-sampling the majority class, highly-absorbed compounds, leads to a balanced distribution (50:50) training set which can achieve better accuracies for poorly-absorbed compounds, whereas the biased training set achieved higher accuracies for highly-absorbed compounds. The use of misclassification costs resulted in improved class predictions, when applied to reduce false positives or false negatives. Moreover, it was shown that the classical overall accuracy measure used in many publications is particularly misleading in the case of unbalanced datasets and more appropriate measures presented here may be used for a more realistic assessment of the classification models' performance. Thus, these strategies offer improvements to cope with unbalanced class datasets to obtain classification models applicable in industry.

Keywords: absorption, QSAR, oral absorption, classification, data distribution, CART, ADME

1. Introduction

Good intestinal absorption is important for oral administration of many drugs in the pharmaceutical industry due to ease of administration and convenience for the patient. The effort to reduce costs and animal testing has resulted in numerous minimisations of assays to become high-throughput[1, 2]. In tandem with these or as an alternative for high-throughput screening assays in early drug discovery, *in silico* modelling using QSAR (Quantitative Structure-Activity Relationships) has been successfully utilised for the prediction of ADMET (absorption, distribution, metabolism, elimination and toxicity) properties, particularly intestinal absorption[3, 4]. QSAR models can act as a tool to filter and highlight undesirable compounds or be used as guidance to help select appropriate assays for the next stage of the drug discovery cascade based on chemical structure and physiochemical properties alone [5, 6].

Datasets in the literature used to predict or classify intestinal absorption are highly biased towards highly-absorbed compounds. This is due to the larger numbers of highly-absorbed compounds amongst the marketed drugs that constitute the datasets[7, 8]. The largest publically available database complied by Hou et al (2007) contains over 80% of compounds with human intestinal absorption values of over 50%. There are a number of reasons for this: the vast majority of percentage human intestinal absorption (%HIA) data is obtained from clinical trials where it is expected for compounds to have good absorption in order to have reached this stage, and the lack of published data representing poorly-moderately absorbed compounds[2, 9, 10]. These biased datasets are not representative of a true industry scenario at present, where there are more drug candidates with poor absorption. With advances of technology, many of the drugs designed today are bigger and more lipophillic, leading to compounds with low absorption due to solubility issues[11]. Imbalanced datasets are a problem for modelling. Any QSAR model produced from a biased dataset will in turn be biased itself to the prediction of the majority class (in this case high absorption class) and will be poorly predictive for the poorly-absorbed compounds.

In order to resolve this, a number of techniques can be carried out to cope with the class imbalance of the dataset. The first technique is to under-sample the majority class (highly-absorbed compounds) in the training set. The problem with this method is the reduction in data utilised for model building, therefore there could be a problem with generalization to new compound sets. This could be resolved by using a bootstrapping technique or bagging[12]. These methods are often used to improve the statistical accuracy and robustness of

predictions, regardless of whether or not the dataset shows class imbalance. However, using a specific version of these methods that under-samples the majority class at all the sampling steps, they may be used to overcome the imbalanced data distribution problem. Ensemble methods such as random forest[13] provide consensus predictions which may have improved accuracy. The problem with bootstrapping and ensemble methods in relation to our work is that they use multiple training sets and therefore multiple decision trees (or classification trees) will be produced, which will increase the complexity and reduce interpretability of the models. Hence, in this work we focus on producing classification models that consist of a single decision tree, to facilitate the interpretability of the model.

Another problem with under-sampling is that in order to assess the predictability of the balanced training set fairly, the validation set will also have to be adjusted to mirror the training set in terms of distribution of the data, but again this reduces the dataset size in the validation set and increases the variability of the results[14]. However the models built using this equal distribution should be better models to predict both poorly and highly-absorbed compounds if a big enough dataset is used.

The second technique applicable to unbalanced class datasets is to increase the cost of misclassification of the minority class. In binary classification there are two types of misclassification, which can be summarised using **Figure 1**.

		Observed class: HIGH	Observed class: LOW
Predicted class	HIGH	True Positive (TP)	False Positive (FP)
	LOW	False Negative (FN)	True Negative (TN)

Figure 1. Possible outcomes of a binary classification

A poorly-absorbed compound misclassified into the highly-absorbed class would be a false positive, and a highly-absorbed compound misclassified into a poorly-absorbed class would be a false negative. Misclassification costs may be defined by the user in the algorithm in order to predict classes that maximise the misclassification cost for false positive or false negatives. An example of this can be shown graphically in **Figure 2**.

		Observed class: HIGH	Observed class: LOW
Predicted class	HIGH	NO COST	2
	LOW	1	NO COST

Figure 2. A classification matrix showing higher misclassification cost assigned to false positives

According to **Figure 2**, if the algorithm attempts to misclassify the poorly-absorbed compound into the highly-absorbed class, there will be a higher cost associated with this misclassification in comparison with the misclassification of a highly-absorbed drug into a poor absorption group. By increasing the cost for misclassification in this example to two the number of false positives should be reduced. The cost assigned to the misclassification can be subjective. However, to assign a number objectively, the class distribution of the high and poorly-absorbed compounds of the training set should be considered by giving a higher cost to the misclassification of poorly-absorbed compounds, the minority class.

In drug discovery there has been a lot of debate on what error to reduce, as both false positives and false negatives have a detrimental effect. The reduction of either one of these errors will depend on the nature of the problem and the intended outcomes, whether this be dependent on business or scientific needs. However, careful consideration on which one to focus on is needed at the start of the project[15]. False negatives give rise to missed opportunities of potential new blockbuster drugs. A potential drug could be dismissed in early library screens as having low absorption when in fact it has high absorption, which is ideal for oral administration[3]. Amlodipine is an example of this, using QSAR it was predicted to have poor bioavailability however it has high observed oral bioavailability[16]. Even if an active compound like Amlodipine is missed and predicted incorrectly, it is less problematic if there are similar compounds in the compound library that are predicted as active and carried through; but problems arise when unique novel chemicals with no similar compounds are missed completely. Despite this, as emphasised by Klopman (2002) care should be taken with the prediction models to avoid overlooking false negatives for the advantage of shortening the drug discovery process [17]. Reducing the number of false positives could be considered

equally as important or more important for cost-effectiveness reasons. If a drug is misclassified as highly-absorbed when in fact it is poorly-absorbed, more time, effort and money is invested to investigate and reveal the compound's true class with further tests. Although there are few publications indicating that false positives need to be decreased rather than false negatives, with the spiralling cost of drug discovery it may be a future consideration for many companies to become more cost-effective. To conclude, although Cummings (2006)[18] states that the trend is to reduce the number of false positives and to put up with the number of false negatives[15], a suitable balance depending on the context of the problem and the intended outcomes may be the answer to reduce time and money testing unsuitable drugs compared with reducing the potential for missed opportunities of new drug candidates, as long as there are still a high number of true positives being discovered[15, 19].

The aim of this work was to use methods specifically designed for coping with unbalanced class datasets in order to improve the classification accuracy of %HIA into high and low classes and finding the best classification model using the Classification & Regression Trees (C&RT) method.

The main dataset used for this work consisted of %HIA data of 645 drugs and drug-like compounds[20]. As stated previously this dataset is biased towards the number of highly-absorbed compounds. In this work two different training sets were randomly selected from this initial dataset, the balanced training set 1 (TS1) contained roughly a 50:50 ratio of highly and poorly-absorbed compounds and the unbalanced training set 2 (TS2) contained roughly an 85:15 ratio of highly and poorly-absorbed compounds respectively.

TS1, having an equal balance of high and poor absorption compounds will be used to show the effects of under-sampling the majority class compared with TS2, the unbalanced dataset. TS1 will also be used to compare the effects of various misclassification cost ratios for reducing either false positives or false negatives. As the TS1 dataset is balanced there is no bias towards reducing either one of these errors, so the effects of misclassification costs will be shown. As shown with the previous discussion regarding which error to reduce, either false positives or false negatives, there is no general consensus, so applying misclassification costs to reduce either error and seeing the results is justified. TS2, containing a higher proportion of highly-absorbed compounds is already biased towards reducing the number of false negatives, so misclassification costs should be assigned to reduce the number of false positives only (by assigning a higher cost to FPs).

2. Methods and Materials

2.1 Dataset

The dataset used consisted of %HIA data for 645 drugs and drug-like compounds extracted from SDF format from the supporting information provided [20]. In this research the 26 compounds containing a quaternary ammonium were removed entirely due to a number of missing molecular descriptors significant to absorption, such as logD, for these compounds; and STATISTICA software would automatically remove compounds with any missing data.[20, 21].

Two training sets and corresponding validation sets were selected from this dataset; training set 1 (TS1) containing roughly a 50:50 ratio, and training set 2 (TS2) containing roughly an 85:15 ratio of highly and poorly-absorbed compounds. The same class distribution for the corresponding validation sets was applied to create a fairer more controlled validation for the models. The exact compound numbers and class distributions are shown in **Table 1**.

Table 1. Compound numbers and class distribution for both training set scenarios

	Number of Compounds		Class Distribution (Ratio of High/Low absorption compounds)	
Dataset	**Training set**	**Validation set**	**Training set**	**Validation set**
TS1	94	89	50:50	50:50
TS2	517	102	85:15	85:15

TS1 is the balanced training set containing about 10 drugs in each 10% range of %HIA. The training set was selected randomly by under-sampling the majority class (highly-absorbed compounds). For the validation set, the remaining compounds were also under-sampled to mimic the data distribution of the training set.

TS2 is the unbalanced training set selected randomly after compounds were sorted by ascending %HIA values and then by logP values. The ascending %HIA values were put into groups of six, then $5/6^{th}$ of these compounds were placed in the training set and the remaining into the validation set. This set is unbalanced as it contains a higher number of the high absorption class.

2.2 Molecular descriptors

A variety of different software packages were used to compute molecular descriptors; they include TSAR 3D v3.3 (Accelrys Inc), MDL QSAR (Accelrys Inc.) and Advanced Chemistry Development ACD Labs/ LogD Suite v12. Due to software limitations some molecular descriptors could not be calculated for some compounds in the dataset. A total of 215 descriptors were used in this study.

2.3 Classification and Regression Trees (C&RT)

STATISTICA v11 (StatSoft Ltd) software was used for classification of compounds using C&RT analysis. According to observed %HIA values in the dataset, compounds were placed into either the 'High' class if %HIA was equal to or greater than 50% or the 'Low' class, if %HIA was less than 50%.

C&RT analysis is a statistical technique that uses decision trees to solve regression and classification problems developed by Breinman et al (1984). If the dependant variable is categorical then a classification tree is made (e.g. predicting low or high absorption classes) and if the dependant variable is continuous then a regression tree is produced resulting in the prediction of numeric %HIA values for all compounds[22, 23].

The binary C&RT analysis starts building the decision tree at the 'tree root' using molecular descriptors. The algorithm in C&RT will choose the most appropriate (statistically significant) molecular descriptor to split the tree and the threshold value to define the split. A parent node splits into two child nodes and then these become the parent groups for the next split. The splitting of the tree continues until it can be no longer split due to stopping factors being applied to prune the tree to prevent over-fitting. The nodes which cannot be split anymore are termed terminal nodes[23, 24], and they contain the predicted classes.

For this work, HIA Class was set as the dependant categorical variable and all 215 molecular descriptors were selected as continuous independent variables. Furthermore, pre-selected subsets of descriptors were used in the analysis. Molecular descriptors were: 1) those chosen by linear stepwise regression and 2) descriptors of Lipinski's rule of five including number of rotatable bonds. Using MINITAB Statistical Software (version 15.1.0.0) linear stepwise regression analysis was performed using the training set TS1 to obtain descriptor sets 1 and 2 and using the training set TS2 to obtain descriptor set 3. During CART analysis, models were created using descriptor sets 1 and 2 for TS1 and descriptor set 3 for TS2. This ensured that the validation set was never used at any stage of model development and remained intact for

the validation of the models. Moreover, Lipinski's 'rule of five descriptors' were used in CART analysis using both TS1 and TS2.

Stopping factors defined in the software will not split a parent node into child nodes if there are less than 10 or 40 compounds in the parent node for TS1 and TS2, respectively. The selection of these stopping factors was based on the statistical performance of the models for the training set as defined by sensitivity (SE), specificity (SP) and SP × SE (defined below). If there was only one compound in a terminal node of a tree, manual pruning was carried out to prevent this final split so that no terminal nodes contained only one compound. In order to cope with missing values in C&RT analysis, STATISTICA can find the next best split variable (molecular descriptor) which is used when the split variable has missing values. The next best split variable ('surrogate') that is chosen is the one that correlates the most with the original one. In this case the optional setting of two surrogates was selected; so if the original variable is missing then the first surrogate variable was used and if this was also missing then the second surrogate variable was used for splitting[25]. If the original variable plus the two surrogates are missing then the compound is removed from the tree. All other settings used were default setting defined by the software.

2.4 Misclassification costs

The aim of a decision tree building algorithm is to create the best model with the lowest total misclassification cost over all compounds in the validation set. By applying varying costs to certain misclassifications (either false positives or false negatives) it is possible to reduce the number of misclassifications due to the higher cost. This study compared the use of the same costing with higher costing to reduce either false positives or false negatives.

For TS1, the balanced dataset, a misclassification cost of two was applied to either reduce false positives or false negatives. As TS2 is unbalanced due to the class distribution of the dataset towards the highly-absorbed compounds (85:15), a misclassification cost ratio of 4:1 was applied to false positive:false negatives. It must be noted that due to the class distributions for TS2 the dataset is already biased towards reducing false negatives, as there are more highly-absorbed compounds than poorly-absorbed.

2.5 Statistical significance of the models

The predictive performance of classification was measured using Specificity (SP), Sensitivity (SE), the cost normalised misclassification index (CNMI) and SP × SE. These terms have been defined below.

Specificity is the ratio of correct classifications of poorly-absorbed compounds (SP = TN/(TN+FP)). In this equation TN is the number of true negatives and FP is the number of false positives. Specificity is inversely proportional to the number of false positives. Sensitivity indicates the correct number of classifications for the highly-absorbed compound class (SE = TP/(TP+FN)), where TP is the number of true positives and FN is the number of false negatives. Sensitivity is inversely proportional to the number of false negatives. Overall accuracy is often defined as the number of correct predictions (true positives and true negatives) divided by the total number of compounds in the validation set. However, this calculation is not suitable to use for this work when the dataset is highly biased. In this case the overall accuracy will be unduly influenced by the classification accuracy of the majority class, the highly absorbed compounds (SE). In this work, in order to represent the overall predictive performance, specificity multiplied by sensitivity was used (SP × SE). SP × SE, although not a very common measure in QSAR, is useful in this case since the data is highly biased.

The cost normalised misclassification index (CNMI) was calculated using Equation 1. The numerator of this equation is calculated by first multiplying the number of each type of misclassifications (false positives and false negatives) by the corresponding misclassification cost and then adding those two products. The denominator (normalization factor) is calculated by first multiplying the total number of compounds in each class – i.e. number of negatives (poorly-absorbed compounds) and number of positives (highly-absorbed compounds) – by the corresponding misclassification costs and then adding those two products.

$$CNMI = \frac{(\mathbf{FP} \times \mathbf{Cost_{FP}}) + (\mathbf{FN} \times \mathbf{Cost_{FN}})}{(\mathbf{Neg} \times \mathbf{Cost_{FP}}) + (\mathbf{Pos} \times \mathbf{Cost_{FN}})} \qquad \textbf{Eq. 1}$$

$Cost_{FP}$ and $Cost_{FN}$ are the misclassification cost assigned for false positives or false negatives and Neg is the total number of poorly-absorbed compounds and Pos is the total number of highly-absorbed compounds. Note that the numerator of **Equation 1** is the total misclassification cost obtained by using a classification model to classify compounds in the validation set, whilst the denominator is the maximum misclassification cost that could in

principle be achieved (if all compounds in the validation set were misclassified). Hence, the calculated value will be between zero and one, with zero representing no misclassification errors, as the number increases to one then the misclassifications of the model increase.

3. Results

Predictive models for classification of drug candidates into high and poor absorption groups are very useful in drug discovery. Unbalanced distribution of data in the available datasets has been a drawback which has traditionally complicated the model development activities. In this work two different training sets with different data distributions and various misclassification costs were used to develop classification trees using the C&RT routine in STATISTICA software. In all result tables the highest SP, SE, SP × SE and the lowest CNMI for the validation sets are highlighted in **bold**. When comparing the models it must be noted that the most significant molecular descriptors selected for splitting the data by the C&RT algorithm will be affected by the class distribution of the training sets, so for TS1 and TS2 with different class distributions different significant descriptors could be picked. Moreover, when comparing models developed using the same training set CNMI maybe a more suitable performance measure since it is normalised for the cost ratios of false positives and false negatives.

3.1. C&RT classification analysis for TS1

Classification using C&RT analysis was carried with the same or different misclassification costs to reduce either false positives or false negatives. Initially all 215 molecular descriptors were set as independent variables and HIA class was set as the dependant categorical variable. In this way C&RT algorithm selects the most significant descriptor out of all 215 for each split. These trees were compared with C&RT trees created by using smaller descriptor sets selected previously by stepwise linear regression using TS1 (descriptor sets 1 and 2), TS2 (descriptor set 3) or descriptors related to Lipinski's rule of five plus number of rotatable bonds[26] (descriptor set 4). The preselected descriptor sets are shown in the **Supporting Information**.

Table 2 shows the predictive performance measures of the classification trees for TS1 obtained with different misclassification costs using all descriptors and descriptor sets 1-4 in the supporting information. Recall that SE, SP and SP × SE measures should be maximized, whilst the CNMI measure should be minimized.

Table 2. The results of C&RT Classification analysis using different descriptor sets and misclassification costs ratios for TS1

Model	Cost FP:FN	Descriptor Set	Set	N validation set	SP × SE	SE	SP	CNMI
1	1:1	ALL	t	83	0.899	0.981	0.917	0.045
			v		0.598	0.733	0.816	0.229
2		1	t	89	0.939	0.962	0.976	0.032
			v		0.625	0.714	0.875	0.213
3		2	t	89	0.951	1.000	0.951	0.021
			v		0.657	0.796	0.825	0.191
4		4	t	89	0.828	0.943	0.878	0.085
			v		0.300	**0.857**	0.350	0.371
5	2:1	ALL	t	83	0.962	0.962	1.000	0.014
			v		0.404	0.667	0.605	0.352
6		1	t	89	0.939	0.962	0.976	0.027
			v		0.547	0.592	**0.925**	0.188
7		2	t	89	0.981	0.981	1.000	0.007
			v		0.604	0.755	0.800	0.203
8		4	t	89	0.920	0.943	0.976	0.034
			v		0.597	0.796	0.750	0.217
9	1:2	ALL	t	83	0.872	0.981	0.889	0.048
			v		0.635	0.778	0.816	0.223
10		1	t	89	0.885	0.981	0.902	0.044
			v		**0.686**	**0.857**	0.800	**0.165**
11		2	t	89	0.951	1.000	0.951	0.015
			v		0.657	0.796	0.825	0.209
12		4	t	89	0.829	1.000	0.829	0.052
			v		0.438	0.796	0.550	0.295

FP = False positive; FN = False negative; SE= Sensitivity, SP = Specificity; CNMI = Cost normalised misclassification index; N validation is the number of validation set compounds that was predicted by the model

In **Table 2,** a cost ratio of 2:1 for FP:FN indicates that a double misclassification cost has been applied for the misclassification of poorly-absorbed compounds compared with the misclassification of highly-absorbed compounds and so forth. Therefore, in this case, the expectation is a reduction in the number of false positives (increased specificity).

In order to see the effect of cost ratios, one should compare the performance measure values of the models generated using the same descriptor set. It can be seen in **Table 2** that when all descriptors were used in the analysis (models 1, 5 and 9) better predictive accuracy is obtained when misclassification costs are adjusted to reduce false negatives (model 9). In this case the SP × SE increased from 0.598 in model 1 to 0.635 in model 9 and the sensitivity was

the highest at 0.778. The CNMI also decreased from 0.229 (model 1) to 0.223 (model 9). This indicates that by applying costs to reduce false negatives a more accurate C&RT model has resulted. The decrease in false negatives (higher sensitivity value) was expected as misclassification costs were to improve the class prediction of highly-absorbed compounds; however the specificity decreased. The classification tree (model 9) has been presented in **Figure 3**. The C&RT trees presented in this paper show on the tree where manual pruning has been carried out. In **Figure 3**, tree manual pruning was not needed.

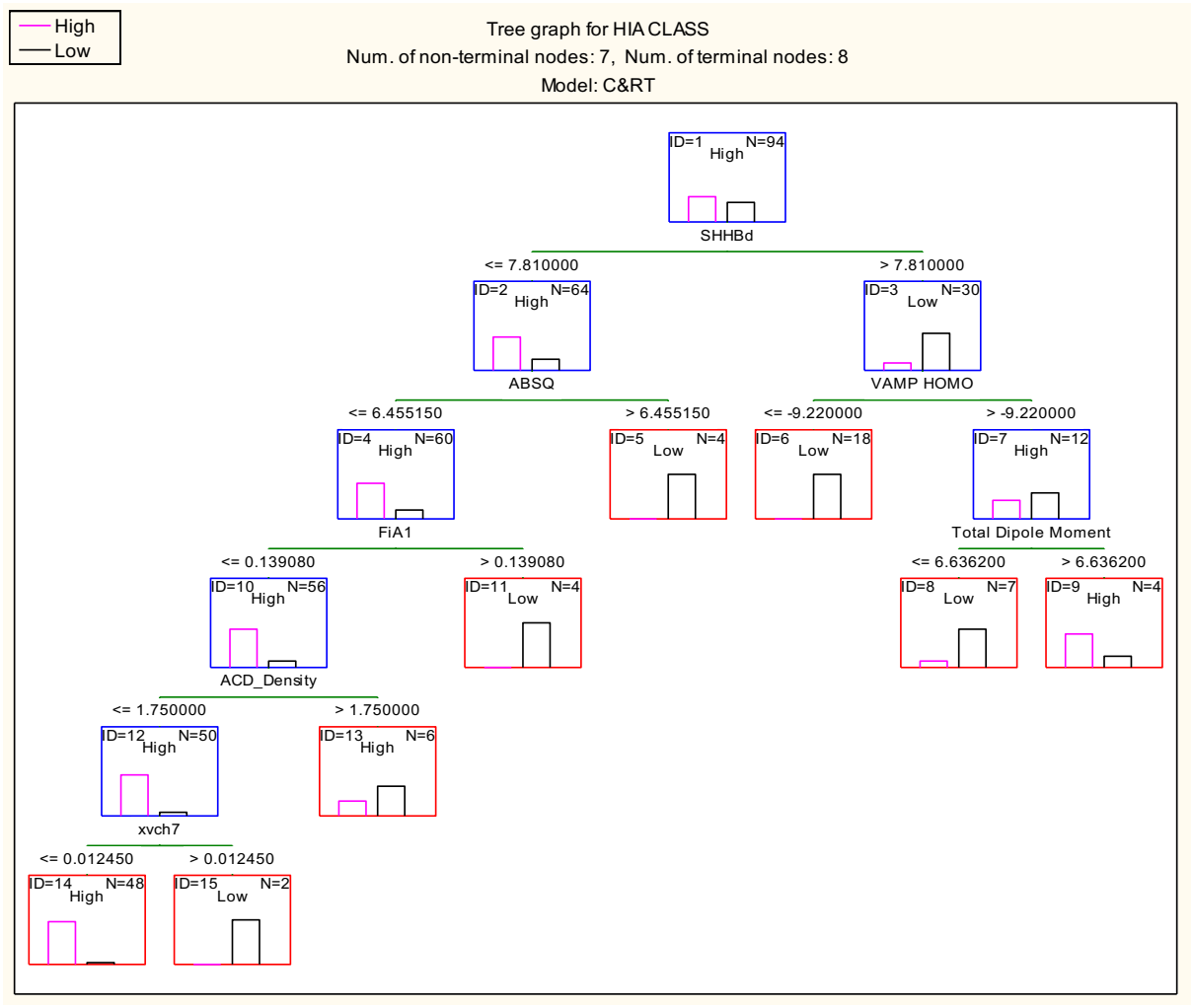


Figure 3. Tree graph for the best C&RT model selecting all molecular descriptors using TS1 training set with misclassification costs applied to reduce false negatives (Model 9)

Furthermore, the molecular descriptors chosen by linear stepwise regression for the estimation of %HIA[26] (descriptors sets 1 and 2) and descriptors of Lipinski's rule of five including number of rotatable bonds (descriptor set 4) were also used in C&RT analysis. **Table 2** shows that the model obtained using descriptor set 1 is the best model (model 10). The fact that most models that are obtained using a pre-selected descriptor sets have better

prediction accuracy indicates that such descriptor selection methods may be better than the descriptor selection algorithm in C&RT. This may be due to the smaller number of chemicals in lower nodes of the tree that are used for the selection of the best descriptor for further splits in C&RT. Model 10 achieved an SP × SE of 0.686, sensitivity value of 0.857 and a specificity value of 0.800 when using a cost ratio of 1:2 for FP:FN. This model has been shown in **Figure 4**. It is interesting to note in **Table 2** that specificity is much better with the model obtained with higher cost for the false positives, using descriptor set 1, which shows that the misclassification costs are having the expected effect on the model.

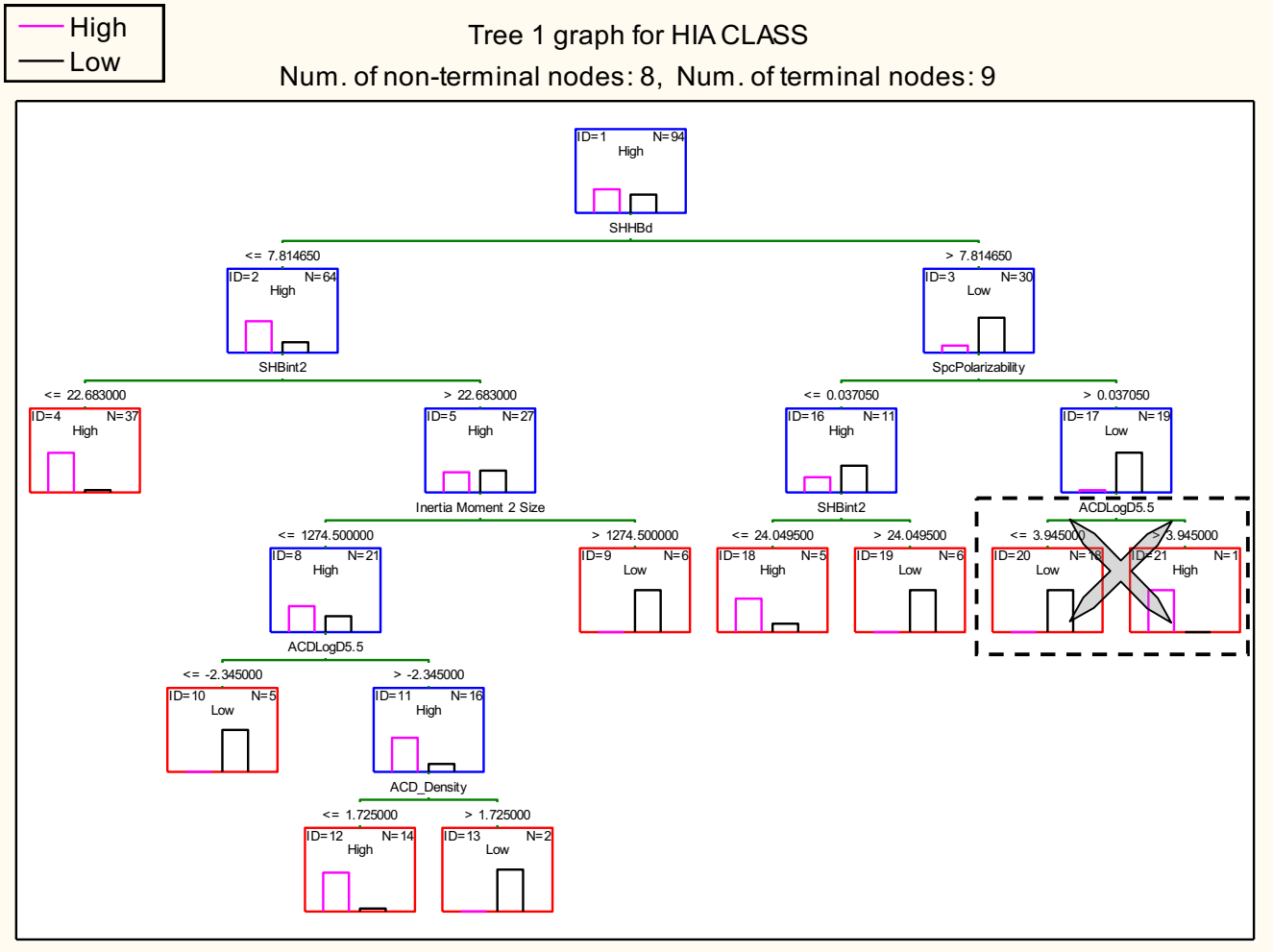


Figure 4. Tree graph for C&RT analysis using TS1 with misclassification costs applied to reduce false negatives using descriptor set 1 (Model 10) - the dashed box around the nodes indicates pruning of the original tree

There is a general pattern when misclassification costs are applied to either reduce false positives or false negatives in the majority of models (**Table 2**). When higher misclassification costs are applied to reduce false positives the specificity values are higher or

equal to models where similar costs are applied, with only a few exceptions. On the other hand, false negative values decrease upon assigning higher misclassification costs on false negatives, resulting in higher or equivalent to FP:FN 1:1 ratio sensitivity values.

Interpretation of the selected models based on TS1

In the tree in **Figure 3** the first split variable is SHHBd, which is the sum of the E-State indexes for hydrogen bond donors. This molecular descriptor is linked to the number of hydrogen bond donors highlighted in Lipinski's rule of five[27], which states that a molecule will be highly likely to be poorly-absorbed if two or more of the following rules are broken: if molecular weight >500 Da, sum of OH and NH hydrogen bond donors >5, calculated logP (ClogP) >5 and sum of N and O atoms as hydrogen bond acceptors >10. The cut off point for SHHBd is 7.81, which corresponds to roughly 3 or more hydrogen bond donor groups. Compounds with low hydrogen bonding donor ability (low SHHBd value) will have poor absorption if ABSQ, the sum of absolute values of atomic partial charges of the molecule[28] is high (node 5). This indicates that molecules or compounds with electronegative or positive atoms (molecules containing heteroatoms) will be less absorbed through the intestine. This is in agreement with the hydrogen bond acceptor factor in Lipinski's rule of five. The compounds with low number of heteroatoms (ABSQ) will have high absorption unless they are highly acidic and have high acidic ionization at pH 1 (FiA1 > 0.139). It has been well cited that drugs that are unionised will pass through the intestinal membrane[27, 29].

The next important descriptor selected by the C&RT for the partitioning of highly hydrogen bond donor compounds is VAMP HOMO, which is the energy of the highest occupied molecular orbital calculated by AM1 semi empirical method and has been used in previous QSAR models for bioavailability[30]. According to this split, compounds with HOMO energy of <= -9.22 are all poorly-absorbed compounds. These are highly polar molecules containing many hydrogen bonding groups (SHHBd) and few or no double bonds – e.g bisphosphonates and macrolides. The high HOMO energy group (Node 7) on the other hand, consists mainly of compounds of moderate absorption level (HIA of 40-60%) and, although marked as highly-absorbed, contains more of the poorly-absorbed compounds to be classified at the next level. These compounds are also of polar nature with many hydrogen bonding groups, but they also have planar areas in the molecule resulting from aromatic groups or other conjugated double bonds (hence high HOMO energy)[31]. High HOMO energy compounds at Node 7 will have high absorption provided that they have dipole moment > 6.63. An

inspection of these compounds at Node 9 shows that these are mainly natural or semi-synthetic compounds e.g. a peptide or a sugar like structure. These compounds may be absorbed by carrier systems due to resemblance to natural metabolites. Examples of these are oxytetracycline, which contains an aromatic system with many oxygen and nitrogen functional groups and is a known substrate of human organic anion transporters[32] or dipyridamole transported via nucleoside transporters in the small intestine[33].

For **Figure 4** (model 10), although all of the eight descriptors of descriptor set 1 were used as independent continuous variables in the C&RT analysis, not all of them were used to build the tree in **Figure 4**; in fact only six out of the eight were used with SHBin7 and SsCH3 not being selected. Similar to model 9, SHHBd is the first split variable in this model. The highly hydrogen bond (according to SHHBd), low absorption group (node 3) has been partitioned again according to SpcPolarizability, which has replaced VAMP HOMO in the previous model (**Figure 3**). SpcPolarizability defines how readily the molecular charge distribution on a molecule, which is the sum of the electronic structure of the individual atoms of the compound, is affected by external oscillating fields. Compounds with low SpcPolarizability values have been divided into groups according to their SHBint2 values. SHBint2 is the sum of E-state indexes for hydrogen bonding groups of path length 2 [34] and is high in compounds like saquinavir and ceftriaxone with peptide bonds. If this value is high then compounds will be classed into the poor absorption class. Compounds with low SHHBd (node 2) have also been partitioned according to SHBint2, with chemicals containing a low number of hydrogen bonding groups of two bond distance showing high oral absorption probability (node 4). Compounds with high SHBint2 may still have high oral absorption if 'inertia moment 2 size' (a size related descriptor) has a low value and ACDlogD5.5 (lipophilicity descriptor) value is high (node 11) and ACD_Density (molecular density) value is small (node 12). Descriptors relating to molecular size have been inversely related to intestinal absorption, therefore the larger the molecule the lower the absorption[35]. The relationship with logD (a measure of hydrophobicity at a specific pH[6]) is in accordance with previous literature[35-38]. ACD Density is the mass per unit volume of a molecule; density will be high for molecules containing many heteroatoms. Compounds with a high density will have low absorption[31], which is also true according to this tree. Pruning of this tree was carried out as there were child nodes with only one compound in them as shown in **Figure 4**.

3.2 C&RT classification analysis for TS2

C&RT classification analysis with misclassification costs was also carried out on TS2, the unbalanced dataset, to see if error rates can be reduced. As there are a larger number of highly-absorbed compounds compared to poorly-absorbed compounds, the misclassification costs to reduce the number of false negatives need not be applied as the class distribution of TS2 already favours the decrease of false negatives. Therefore misclassification costs are applied for reducing false positives only. The costing of 4 was applied to false positives (keeping the baseline cost of 1 for false negatives), as this was considered the most suitable number based on the class distribution of roughly 4:1 for high to low absorption compounds. The results of the C&RT classification analysis for TS2 are shown in **Table 3**.

Table 3. The results of C&RT Classification analysis using different descriptor sets and misclassification costs ratios for TS2

Model	Cost FP:FN	Descriptor Set	Set	N Validation set	SP × SE	SE	SP	CNMI
13	1:1	ALL	t	94	0.862	0.955	0.903	0.053
			v		0.400	0.880	0.455	0.170
14	1:1	3	t	102	0.704	0.973	0.724	0.064
			v		0.445	0.954	0.467	0.118
15	1:1	4	t	102	0.620	0.982	0.632	0.070
			v		0.451	**0.966**	0.467	0.108
16	4:1	ALL	t	94	0.861	0.873	0.986	0.033
			v		**0.660**	0.807	**0.818**	**0.070**
17	4:1	3	t	102	0.879	0.890	0.987	0.028
			v		0.653	0.816	0.800	0.077
18	4:1	4	t	102	0.855	0.890	0.961	0.033
			v		0.517	0.862	0.600	0.099

FP = False positive; FN = False negative; SE= Sensitivity, SP = Specificity; CNMI = Cost normalised misclassification index; N validation is the number of validation set compounds that was predicted by the model

Table 3 shows that when all descriptors were available to C&RT analysis the best results were achieved when applying misclassification costs to reduce false positives (comparing model 13 and 16). As expected, specificity increases and misclassification error rate decreases when misclassification costs were applied. By applying misclassification costs to increase specificity, the sensitivity of the model will decrease (**Table 3**). **Figure 5** shows the best model when all descriptors were supplied and the significant descriptors were selected by C&RT analysis (model 16).

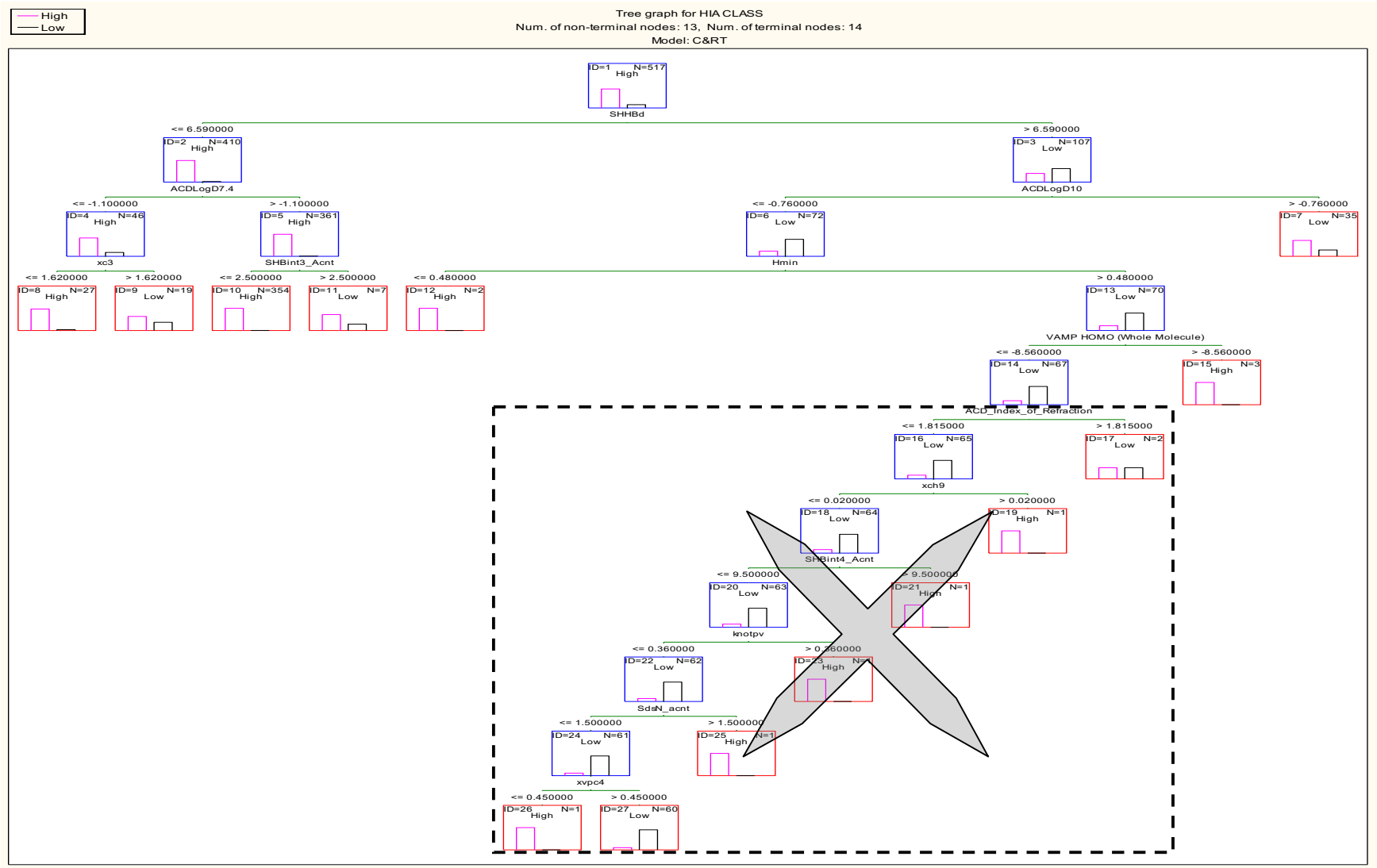


Figure 5. Tree graph for the best C&RT analysis using TS2 using all descriptors with misclassification costs applied to reduce false positives (Model 16) - the dashed box around the nodes denotes pruning of the original tree

As with the TS1, C&RT analysis was carried out using the pre-selected molecular descriptors[26] (**Table 2**). **Table 3** shows that the best pre-selected descriptor set was descriptor set 3 (model 17) when considering SP × SE. The classification tree model 17 is shown in **Figure 6**. With misclassification costs to reduce false positives the tree had the highest specificity (0.800) and also the lowest CNMI (0.077). Depending on the use of the model, if the reduction of false positives (increase of specificity) is the intention then using misclassification costs will increase the specificity for descriptor set 3 from 0.467 to 0.800; however, sensitivity decreases from 0.954 to 0.816.

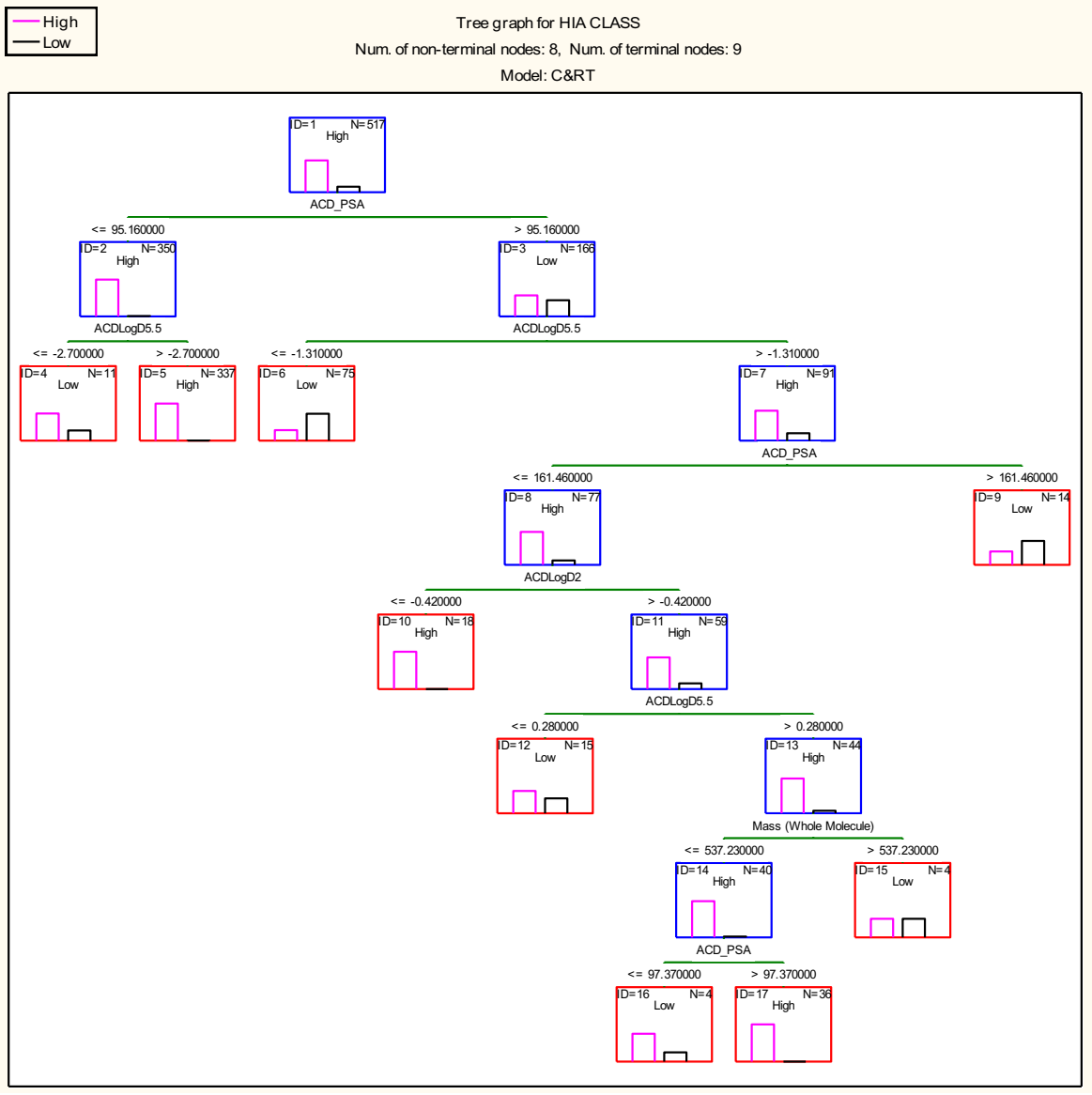


Figure 6. Tree graph for C&RT analysis using TS2 with misclassification costs applied to false positives (FP:FN 4:1) using descriptor set 3 (Model 17)

Interpretation of selected models based on TS2

Figure 5 shows the selected tree when C&RT analysis selected the descriptors from all the supplied descriptors (model 16). Similar to models 9 and 10 obtained using TS1, this tree involves the hydrogen bond donor descriptor, SHHBd, as the first variable. Compounds with high SHHBd values are more likely to have poor oral absorption, especially if they are hydrophilic with ACDLogD10 below -0.76; unless their Hmin value is lower than 0.48. A high number of potential H-bond formations is detrimental to high oral absorption, which is cited in the literature[11, 27, 39, 40]. Hmin is the minimum hydrogen electrotopological-state value for all atoms in the drug molecule and shows the nature of the hydrogen atoms attached to the skeleton of the drug molecule and whether they are hydrogen bond donors[41]. Otherwise if the Hmin value is higher than 0.48 compounds with higher VAMP HOMO than -8.56 may still have high oral absorption, but the large majority of compounds have a lower HOMO energy value and therefore will be expected to be poorly-absorbed through the gastrointestinal system (node 14). On the left hand side of the tree (node 2), for compounds with low hydrogen bond donor ability (SHHBd ≤ 6.59), oral absorption is expected to be high, unless ACDlogD7.4 is low and xc3 is high (node 9). The descriptor xc3 is the third order cluster chi connectivity index. This Chi index encodes the number and branching of the molecule for a single branch point and in this tree, it indicates that branched molecules (of hydrophilic nature) have poor oral absorption[42]. It must be noted in Figure 5 in nodes 9 and 11, for example, that the effect of misclassification costs is altering the final terminal class node, showing the misclassification costs applied to reduce false positives is working. Moreover, for high ACDlog7.4 compounds (>-1.10) oral absorption would be poor if they have a high number of internal hydrogen bonding groups of three bond distance (SHBint3_Acnt). It is interesting to note those nodes in the tree after the first split using SHHBd were both logD molecular descriptors but at different pH values. LogD at different pH values is affected by the ionization of the compound and is related to the compound's pka. For example, for logD10, which means the diffusion coefficient at pH10, any basic compounds at pH10 will be unionized, therefore will have higher logD10 values than acidic compounds which will remain ionized due to the higher pH, and in the case of intestinal absorption will then be not absorbed. This indicates that pH-dependent lipophilicity measure (logD) at different pH values are important in distinguishing between high and low absorption for acidic and basic compounds as well as characterizing the lipophilicity.

In **Figure 6**, all the molecular descriptors have been described previously apart from polar surface area (PSA). This descriptor along with lipophilicity has been described as an

influential molecular descriptor in predicting passive intestinal absorption[6, 43, 44]. PSA is the area of the Van der Waals surface that arises from oxygen and nitrogen atoms or hydrogen atoms bound to these atoms. It is a polarity measure which is also related to size and has a negative correlation with intestinal absorption so the larger the PSA the lower the absorption. Mass has also been used in this tree and in accordance with Lipinski's rule of five, but only with a slightly different cut off point of 537.23 Da[27, 45].

4. Discussion

4.1 Comparing models

There are many statistical measures for the assessment of the predictability of classification models. The most common in QSAR literature are overall accuracy accompanied by SP and SE. However it must be emphasized here that 'accuracy' reported in the literature as the ratio of all the correctly classified compounds is misleading when the datasets are highly skewed[22, 46, 47]. In other words, due to the majority of highly-absorbed compounds in the training and validation sets, the classification outcome of these compounds disproportionately affects the overall accuracy: therefore accuracy will follow the same trend as the sensitivity values in the model and fail to take into account the specificity appropriately. For example, if a dataset contained 90% of highly-absorbed and 10% of poorly absorbed compounds, a trivial classifier would consist of predicting the highly-absorbed class (the majority class) for all compounds in the validation set. Such a trivial majority classifier – which does not involve any data analysis – would trivially achieve an overall accuracy of 90% (if accuracy is simply measured as (TP + TN) / (TP + TN + FP + FN)). However, this high accuracy is misleading. Although the majority classifier achieved perfect prediction for the high absorption class, it achieved no correct predictions for the poorly –absorbed class. This example clearly shows a weakness of the overall accuracy measure, which is not an appropriate measure to use when the class distribution is very unbalanced. The use of SP × SE avoids the above problem in this scenario, since the trivial majority classifier would achieve a prediction of 0% by multiplying the sensitivity (100%) and specificity (0%), and therefore would show the majority classifier's ability to classify both classes as poor. A measure of 0% accuracy for the majority classifier is also intuitively fair; since that classifier is not even taking a look at the value of the descriptors (it just counts the number of compounds in each class). In summary, the overall accuracy as defined in the literature is an incorrect measure of accuracy in problems with very unbalanced class distributions, like the dataset in this work. In this work, to overcome the problem, the better accuracy measure of SP × SE was used as the measure of

the overall accuracy. This measure is a better representation of a models' predictability, as it is affected the same way by false negatives as by false positives, and should be used in measuring the overall accuracy of a classification model. Other measures such as the Matthew's correlation coefficient (MCC)[48], kappa statistic[49] and Youden's J statistic[50] to name a few have been used in the literature, with the choice of which one to use being subjective, with each measure having advantages and disadvantages[51, 52]. Youden's index or J statistic[50] is used frequently for medical diagnosis tests and is defined as 1 – (SE + SP). Kappa has been introduced as a chance-corrected measure of accuracy[49] but it uses the overall accuracy in the calculation, which may not be suitable in this case since it will have a higher contribution from the majority highly absorbed class. MCC is another useful measure frequently used in QSAR and although it uses all four numbers (TP, TN, FP, FN), it requires normalized distribution and may give controversial results, for example when there are very few FP but also there are very few TPs[52]. We used SP × SE on the grounds that it is a simple measure with a clear interpretation and gives an overall fair measure of model performance without being affected by the class distribution bias.

As stated previously, a direct comparison between the two different training sets is not a fair comparison due to the different class distributions of TS1, the balanced set, and TS2, the set biased towards highly-absorbed compounds. Nevertheless, it can be seen that TS1 in the majority of cases leads to higher specificity when misclassification costs are equal for FN and FP. TS2 gave higher sensitivity in all cases, which is expected due to the bias of the training set towards highly-absorbed compounds. It has been cited that the rule of five can give rise to false positives and could be a possible explanation why the specificity is lower for this model even with a balanced training set[27, 53, 54].

When misclassification costs are applied to either TS1 or TS2 to reduce false positives, specificity improves for both training sets. Moreover, it can be seen in **Table 3** that the use of 4:1 misclassification costs for FP:FN leads to improved models for TS2 with better SP × SE values. This finding shows that using misclassification costs can overcome a dataset bias by increasing specificity.

In this paper we compared the effect of allowing the software to pick the most significant descriptors from all 215 descriptors used or from a smaller subset of descriptors previously selected as significant by stepwise regression analysis or those related to Lipinski's rule of five. **Table 2** shows that model 10 achieved the lowest value CNMI of 0.165 and the highest

SP × SE of 0.686 using descriptor set 1. The next best CNMI was again achieved by descriptor set 1 with a value of 0.188 (model 6); this model also obtained the highest specificity of 0.925 when misclassification costs were applied to reduce false positives. From **Table 2** it is interesting to see that in several cases the CNMI values are higher for C&RT models using all descriptors compared with those models using smaller descriptor sets selected by feature selection techniques, meaning that there are more misclassification errors when allowing the C&RT analysis to pick significant descriptors from the 215 available. This could show that using linear stepwise regression to select a smaller subset of significantly relevant molecular descriptors to intestinal absorption beforehand can be advantageous as often models are produced with fewer misclassifications. The most accurate models for TS2 (**Table 3**) are models 16 and then 17 which was developed using all descriptors or descriptor set 3. The fact that using all descriptors works well for TS2 but for TS1 prior descriptor selection is best, suggests that C&RT can be an efficient descriptor selection method when a large dataset is used (517 vs 94 compounds in TS2 and TS1, respectively).

In the study for TS1 containing 94 compounds, a validation set containing 89 compounds has been used which mirrors the balanced data distribution of the training set. By balancing the validation set too it gives a fair representation of the models' predictive performance. As an additional test the predictive performance of the models was investigated for a new validation set containing all the compounds not used in the training set. It must be noted that the additional validation set compounds are all highly-absorbed with the exception of two compounds. Therefore this validation set is biased. The results of this work can be found in **Table 4**.

Table 4. The validation results of C&RT Classification models obtained using TS1 for all the remaining compounds not used in training

Model	Cost FP:FN	Descriptor Set	N validation set	SP x SE	SE	SP	CNMI
1	1:1	ALL	496	0.647	0.869	0.745	0.143
2		1	521	**0.730**	0.852	0.857	0.148
3		2	521	0.728	0.887	0.820	0.119
4		4	521	0.341	0.898	0.380	0.152
5	2:1	ALL	496	0.510	0.800	0.638	0.131
6		1	521	0.712	0.775	**0.918**	0.115
7		2	521	0.685	0.856	0.800	0.089

8		4	521	0.654	**0.909**	0.720	**0.072**
9	1:2	ALL	496	0.673	0.855	0.787	0.258
10		1	521	0.701	0.881	0.796	0.214
11		2	521	0.657	0.887	0.740	0.208
12		4	521	0.497	0.887	0.560	0.224

FP = False positive; FN = False negative; SE= Sensitivity, SP = Specificity; CNMI = Cost normalised misclassification index; N validation is the number of validation set compounds that was predicted by the model

According to **Table 4** the best models according to SP × SE were those using descriptor set 1 (models 2, 6 and 10), which corresponds to the results seen earlier for the smaller balanced validation set (**Table 2** Model 10).

4.2 Discussion of the related literature

Summary tables in the literature detail the accuracy, specificity and sensitivity of classification work carried out by previous studies[55-57]. In particular Talevi et al (2011) has compiled a summary table summarising classification studies of intestinal absorption over the past decade. To compare the models obtained in this work with the literature is a very difficult task. There is lack of compound information and data distribution and a lack of consistency in the literature with regards to validation techniques and more importantly how the overall accuracy of the models are measured[58-60]. To directly compare our work with others in the literature all the information as described previously would be needed to mimic conditions regarding the dataset to enable comparison of the models, however this is not freely available[61].

The number of compounds in the datasets in the literature should be considered when assessing the model performances. Small datasets may achieve high prediction accuracy within the chemical space of the dataset, however will lack generalization to new chemical compounds. In the study by Niwa et al (2003), using 67 compounds achieved 100% correct classification for the training set, however this dropped to 80% for the external prediction set of 12 compounds. It must be highlighted that the main misclassification in Niwa et al's model was for the poorly-absorbed compounds, which were represented inadequately in Niwa's dataset[47]. As a result, the overall accuracy of Niwa et al's model as calculated using our accuracy measurement (SP × SE) yields a value of 0.667. This is a reoccurring problem with the other datasets in the literature that we considered[7, 8, 17]. Poorly-absorbed compounds are predicted better using our models due to the larger representation of this class in our TS1 training set and/or the use of varying misclassification costs.

Perez et al (2004)[62] used linear discriminant analysis to classify a dataset of 209 compounds with training and validation set of 82 and 127 compounds respectively. This paper created two models, one which focussed on classification of %HIA using a threshold of ≤30% HIA and the other focussing on classification using a threshold of >80% HIA. Both training and validation sets are heavily biased towards highly-absorbed compounds and the results reflect this. Higher sensitivity values of 0.955 and 0.835 and much lower specificity values of 0.765 and 0.722 for the threshold of ≤30% and >80% respectively.

As stated before, in most studies the accuracy and sensitivity results are higher than the specificity values, due to the under-representation of poorly-absorbed compounds[22, 46, 47]. The one exception to this was obtained by Hou et al (2007) who obtained higher specificity than sensitivity in the validation set. However only five compounds in their validation set was defined as poorly-absorbed. Deconinck et al (2005) carried out C&RT analysis using Splus software on a smaller subset of the dataset compiled by Hou et al (2007). Deconinck achieved, using C&RT as a variable selection classification technique using a validation set of 27 compounds, an overall prediction accuracy of 85%. However this validation set only contained highly-absorbed compounds and therefore only sensitivity values could be considered[22].

Lipinski's rule of five is a qualitative rule based model which indicates that poor absorption is highly likely when two or more of the rules are broken. It has been criticised for having a high rate of false positives[53, 54]. With this work, descriptors describing Lipinski's rule of five plus the number of rotatable bonds allowed a qualitative evaluation of Lipinski's rule of five via C&RT analysis. Using Lipinski's rule of five in its original form (if 2 or more rules were violated indicating poor absorption) specificity was 0.425 and 0.400 for the validation sets of TS1 and TS2 respectively. By incorporating these descriptors (descriptor set 4) in C&RT, for TS1 upon using higher misclassification costs to reduce false positives the specificity was 0.750 and for TS2 specificity was 0.600. Using misclassification costs to reduce false positives, an improvement to Lipinski's rules using misclassification costs was made possible.

The descriptors selected in the models can be interpreted according to the known mechanisms involved in the absorption process. **Table 5** gives a summary of all the molecular descriptors used in the selected models of 9, 10, 16 and 17. The most common molecular descriptors

used in the best models were descriptors of hydrogen bonding (such as SHHBd, SHBint2), log D at various pH values which is related to lipophilicity and acid/base property, ACD_Density which is related to the number of heteroatoms in the molecules, and polar surface area (PSA) which has been cited as a molecular descriptor relating to polarity and size [20, 63]. Other important molecular attributes are size related parameters. These are in agreement with the literature indicating that the molecular descriptors important to intestinal absorption are those related to lipophilicity, hydrogen bonding, polarity, ionization, and size[45, 60, 63]. From considering the molecular descriptors utilised in the models in this work overall, no matter what training set used, molecular descriptors that described these parameters which are shown in the literature to be important for intestinal absorption were present in the best models[30, 64].

Table 5. Molecular descriptors used in the selected models (9, 10, 16 and 17)

Type of descriptor	Name of descriptor	Number of occurrences in selected models
Hydrogen bonding	SHHBd	3
	ABSQ	1
	ACD_Density	2
	SHBint2	1*
	SHBint3_Acnt	1
	Hmin	1
	ACD_PSA	1*
Lipophilicity	ACD logD 5.5	2*
	ACD logD 7.4	1
	ACD logD 10	1
	ACD logD 2	1
Size	xvch7	1
	Inertia moment 2 (size)	1
	xc3	1
	Mass	1
Polarity/ polarizability	VAMP HOMO	2
	Total dipole moment	1
	Spc polarizability	1
Acidity	FiA1	1

* occurred more than once in a single tree model

5 Conclusion

Class imbalance occurs frequently in QSAR and drug discovery datasets[14, 65-67]. This could be for a number of reasons; however in this context it is due to lack of publically available data for the minority class, poorly-moderately absorbed compounds, in the literature. The aim of this work was to improve the class prediction of poorly absorbed compounds by the use of varying misclassification costs in C&RT analysis. This was analysed using two training sets, the one selected by under-sampling the majority class (TS1), or the training set selected randomly and hence biased towards highly-absorbed compounds (TS2). The comparison between C&RT descriptor selection and pre-selecting a small subset of molecular descriptors using statistical techniques or rule-based models was also considered. In this work, in order to effectively compare the models, the traditional 'overall accuracy' measure was scrutinised and better measures of prediction accuracy, SP × SE, and cost normalised misclassification index (CNMI), were suggested and incorporated.

Under-sampling the majority class to create a balanced training set produced models that had high predictive power for the prediction of poorly-absorbed compounds. The randomly selected training set (TS2) as expected had high predictive power for highly-absorbed compounds with high sensitivity values, but this was accompanied by low specificity values. This conclusion conforms to the previous work using regression and discriminant analysis classification[26].

The use of misclassification costs led to improvements in prediction accuracy. Even though there is no general consensus to reduce false positives or false negatives from the literature, this work shows that misclassification costs can be applied to reduce false positives or false negatives. Other considerations such as poor solubility and carrier mediated transport systems can play a part in misclassification error rates in the models[17]. For the biased training set containing the majority high absorption class, applying higher costs for the misclassification of false positives improved specificity in all cases. The imbalanced dataset can be utilised without removing compounds as an advantage for improved sensitivity as it will already be biased towards high absorption compounds. Therefore, varying ratios of misclassification costs can be used as a vital and effective tool to overcome class imbalance, which is a recurring problem in drug discovery datasets.

The comparison between using all descriptors for the C&RT or to use a smaller subset of molecular descriptors suggests that the descriptors selected by stepwise linear regression may

achieve better prediction, but this cannot be generalized and descriptor selection by C&RT may work just as well when a large training set is used, e.g. TS2.

In conclusion, reasonably interpretable user friendly models that can be easily understood and utilised for specific purposes has been achieved by using two strategies, under-sampling the majority class of the training set and misclassification costs, to overcome class imbalance of the dataset.

Supporting Information

The molecular descriptor sets selected using stepwise regression analysis for different training sets (descriptor sets 1-3) and molecular descriptors used in Lipinski's rule of five plus number of rotatable bonds (descriptor set 4) is described in the Supporting Information. This information is available free of charge via the Internet at http://pubs.acs.org.

References

1. Davis, A. M.; Keeling, D. J.; Steele, J.; Tomkinson, N. P.; Tinker, A. C. Components of successful lead generation. *Curr. Top. Med. Chem.* **2005**, *5*, 421-439.

2. Gleeson, M. P.; Hersey, A.; Hannongbua, S. In-Silico ADME Models: A General Assessment of their Utility in Drug Discovery Applications. *Curr. Top. Med. Chem.* **2011**, *11*, 358-381.

3. Yu, H. S.; Adedoyin, A. ADME-Tox in drug discovery: integration of experimental and computational technologies. Drug Discovery Today. **2003**, *8*, 852-861.

4. Chohan, K. K.; Paine, S. W.; Waters, N. J. Advancements in Predictive In Silico Models for ADME. Curr. Chem. Biol. **2008**, *2*, 215-228.

5. Geerts, T.; Heyden, Y. V. In Silico Predictions of ADME-Tox Properties: Drug Absorption. *Comb. Chem. High Throughput Screening*. **2011**, *14*, 339-361.

6. van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery*. **2003**, *2*, 192-204.

7. Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726-735.

8. Zhao, Y. H.; Abraham, M. H.; Le, J.; Hersey, A.; Luscombe, C. N.; Beck, G.; Sherborne, B.; Cooper, I. Rate-limited steps of human oral absorption and QSAR studies. *Pharm. Res.* **2002**, *19*, 1446-1457.

9. Oprea, T. I.; Allu, T. K.; Fara, D. C.; Rad, R. F.; Ostopovici, L.; Bologa, C. G. Lead-like, drug-like or "pub-like": how different are they? *J. Comput.-Aided Mol. Des.* **2007**, *21*, 113-119.

10. Yan, A.; Wang, Z.; Cai, Z. Prediction of Human Intestinal Absorption by GA Feature Selection and Support Vector Machine Regression. *Int. J. Mol. Sci.* **2008**, *9*, 1961-1976.

11. Thomas, V. H.; Bhattachar, S.; Hitchingham, L.; Zocharski, P.; Naath, M.; Surendran, N.; Stoner, C. L.; El-Kattan, A. The road map to oral bioavailability: an industrial perspective. *Expert Opin. Drug Metab. Toxicol*. **2006**, *2*, 591-608.

12. Breiman, L., Bagging predictors. *Mach. Learn.* **1996**, *24*, 123-140.

13. Breiman, L., Random forests. *Mach. Learn.* **2001**, *45*, 5-32.

14. Blagus, R.; Lusa, L. Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics.* **2010**, *11*.

15. White, R. E. High-throughput screening in drug metabolism and pharmacokinetic support of drug discovery. *Annu. Rev. Pharmacol. Toxicol.* **2000**, *40*, 133-157.

16. Beresford, A. P.; Segall, M.; Tarbit, M. H. In silico prediction of ADME properties: Are we making progress? *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 36-42.

17. Klopman, G.; Stefan, L. R.; Saiakhov, R. D. ADME evaluation 2. A computer model for the prediction of intestinal absorption in humans. *Eur. J. Pharm. Sci.* **2002**, *17*, 253-263.

18. Cummings, D., J. Pharmaceutical Drug Discovery: Designing the Blockbuster Drug. In *Screening Methods for Experimentation in Industry, Drug Discovery, and Genetics*, First edition; Dean, A.; Lewis, S., Eds. Springer: New York, **2006**; pp 74-76.

19. Rydzewski, M. R. *Real World Drug Discovery A Chemist's Guide to Biotech and Pharmaceutical Research*; First edition; Elsevier: Oxford, **2008**.

20. Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model*. **2007**, *47*, 208-218.

21. Abraham, M. H.; Zhao, Y. H.; Le, J.; Hersey, A.; Luscombe, C. N.; Reynolds, D. P.; Beck, G.; Sherborne, B.; Cooper, I. On the mechanism of human intestinal absorption. *Eur. J. Med. Chem.* **2002**, *37*, 595-605.

22. Deconinck, E.; Hancock, T.; Coomans, D.; Massart, D. L.; Vander Heyden, Y. Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *J. Pharm. Biomed. Anal.* **2005**, *39*, 91-103.

23. Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees.* First edition; Chapman and Hall/CRC: Boca Raton, **1984**.

24. Tan, P. N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*. First edition; Pearson International Edition: Boston, **2006**.

25. Witten, I. H.; Frank, E.; Hall, M. A. *Data Mining Practical Machine Learning Tools and Techniques*. Third edition; Morgan Kaufmann Publishers: Burlington, **2011**.

26. Ghafourian, T.; Newby, D.; Frietas, A. A. The impact of training set data distributions for modelling of passive intestinal absorption. *Int. J. Pharm*. **2012**, *436*, 711-720.

27. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev*. **1997**, *23*, 3-25.

28. Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron*. **1980**, *36*, 3219-3228.

29. Pang, K. S. Modeling of intestinal drug absorption: Roles of transporters and metabolic enzymes (for the Gillette Review Series). *Drug Metab. Dispos*. **2003**, *31*, 1507-1519.

30. Turner, J. V.; Glass, B. D.; Agatonovic-Kustrin, S. Prediction of drug bioavailability based on molecular structure. *Anal. Chim. Acta.* **2003**, *485*, 89-102.

31. Agatonovic-Kustrin, S.; Beresford, R.; Yusof, A. P. M. Theoretically-derived molecular descriptors important in human intestinal absorption. *J. Pharm. Biomed. Anal.* **2001**, *25*, 227-237.

32. Sai, Y.; Tsuji, A. Transporter-mediated drug delivery: recent progress and experimental approaches. *Drug Discovery Today*. **2004**, *9*, 712-720.

33. Lin, W. W.; Buolamwini, J. K. Synthesis, flow cytometric evaluation, and identification of highly potent dipyridamole analogues as equilibrative nucleoside transporter 1 inhibitors. *J. Med. Chem.* **2007**, *50*, 3906-3920.

34. Wanchana, S.; Yamashita, F.; Hara, H.; Fujiwara, S. I.; Akamatsu, M.; Hashida, M. Two- and three-dimensional QSAR of carrier-mediated transport of beta-lactam antibiotics in Caco-2 cells. *J. Pharm. Sci.* **2004**, *93*, 3057-3065.

35. Varma, M. V. S.; Obach, R. S.; Rotter, C.; Miller, H. R.; Chang, G.; Steyn, S. J.; El-Kattan, A.; Troutman, M. D. Physicochemical Space for Optimum Oral Bioavailability: Contribution of Human Intestinal Absorption and First-Pass Elimination. *J. Med. Chem.* **2010**, *53*, 1098-1108.

36. Zakeri-Milani, P.; Tajerzadeh, H.; Islambolchilar, Z.; Barzegar, S.; Valizadeh, H. The relation between molecular properties of drugs and their transport across the intestinal membrane. *Daru, J. Pharm. Sci.* **2006**, *14*, 164-171.

37. Comer, J. E. A. High-throughput Measurement of log D and pka. In *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability (Methods and Principles in Medicinal Chemistry)*, First Edition; van de Waterbeemd,; Lennernäs, H.; Artursson, P.; Mannhold, R.; Kubinyi, H.; Folkers, G. Eds. Wiley-VCH: Weinheim, **2003**; Vol. 18, p 23.

38. Kerns, E. H.; Di, L. *Drug like properties: Concepts, Structure Design and Methods from ADME to Toxicity Optimisation*. First edition; Academic Press Elsevier: Burlington, **2008**.

39. Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods.* **2000**, *44*, 235-249.

40. Yu, K.; Chen, F.; Li, C. Absorption, Disposition, and Pharmacokinetics of Saponins from Chinese Medicinal Herbs: What Do We Know and What Do We Need to Know More? *Curr. Drug Metab*. **2012**, *13*, 577-598.

41. Wang, Y.; Li, Y.; Ding, J.; Jiang, Z.; Chang, Y. Estimation of bioconcentration factors using molecular electro-topological state and flexibility. *SAR QSAR Environ. Res*. **2008**, *19*, 375-395.

42. Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure-Property Modeling. In *Reviews in Computational Chemistry*, Boyd, D.; Lipkowitz, K., Eds. VCH: New York, **1991**; pp 384–385.

43. Palm, K.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **1996**, *85*, 32-39.

44. van de Waterbeemd, H.; Kansy, M. Hydrogen-Bonding Capacity and Brain penetration. *Chimia* **1992**, *46*, 299-303.

45. Yang, Y. D.; Engkvist, O.; Llinas, A.; Chen, H. M. Beyond Size, Ionization State, and Lipophilicity: Influence of Molecular Topology on Absorption, Distribution, Metabolism, Excretion, and Toxicity for Druglike Compounds. *J. Med. Chem*. **2012**, *55*, 3667-3677.

46. Deconinck, E.; Zhang, M. H. H.; Coomans, D.; Vander Heyden, Y. Classification tree models for the prediction of blood-brain barrier passage of drugs. *J. Chem. Inf. Model*. **2006**, *46*, 1410-1419.

47. Niwa, T. Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 113-119.

48. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **1975**, *405*, 442-451.

49. Cohen, J. Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull*. **1968**, *70*, 213-220.

50. Youden, W. J. Index for rating diagnostic tests. *Cancer*. **1950**, *3*, 32-35.

51. Gleeson, M. P.; Modi, S.; Bender, A.; Robinson, R. L. M.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; Glen, R. C. The Challenges Involved in Modeling Toxicity Data In Silico: A Review. *Curr. Pharm. Des.* **2012**, *18*, 1266-1291.

52. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. **2000**, *16*, 412-424.

53. Andrews, C. W.; Bennett, L.; Yu, L. X. Predicting human oral bioavailability of a compound: Development of a novel quantitative structure-bioavailability relationship. *Pharm. Res*. **2000**, *17*, 639-644.

54. Zhu, J. Y.; Wang, J. M.; Yu, H. D.; Li, Y. Y.; Hou, T. J. Recent Developments of In Silico Predictions of Oral Bioavailability. *Comb. Chem. High Throughput Screening*. **2011**, *14*, 362-374.

55. Suenderhauf, C.; Hammann, F.; Maunz, A.; Helma, C.; Huwyler, J. Combinatorial QSAR Modeling of Human Intestinal Absorption. *Mol. Pharmaceutics*. **2011**, *8*, 213-224.

56. Hou, T. J.; Wang, J. M.; Zhang, W.; Wang, W.; Xu, X. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.* **2006**, *13*, 2653-2667.

57. Talevi, A.; Goodarzi, M.; Ortiz, E. V.; Duchowicz, P. R.; Bellera, C. L.; Pesce, G.; Castro, E. A.; Bruno-Blanch, L. E. Prediction of drug intestinal absorption by new linear and non-linear QSPR. Eur. J. Med. Chem. **2011**, *46*, 218-228.

58. Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X. Q.; Doweyko, A.; Li, Y. In silico ADME/Tox: why models fail. *J. Comput.-Aided Mol. Des*. **2003**, *17*, 83-92.

59. The, H. P.; Gonzalez-Alvarez, I.; Bermejo, M.; Sanjuan, V. M.; Centelles, I.; Garrigues, T. M.; Cabrera-Perez, M. A. In Silico Prediction of Caco-2 Cell Permeability by a Classification QSAR Approach. *Mol. Inf*. **2011**, *30*, 376-385.

60. Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2001**, *90*, 749-784.

61. Davis, A. M.; Brunea, P. In Silico Prediction of Solubility In *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability (Methods and Principles in Medicinal Chemistry)*, First edition.; Van de Waterbeemd, H.; Lennernäs, H.; Artursson, P.; Mannhold, R.; Kubinyi, H.; Folkers, G., Eds. Wiley-VCH: Weinheim, **2003**; Vol. 18, pp 53-56.

62. Perez, P. A. C.; Sanz, M. B.; Torres, L. R.; Avalos, R. C.; Gonzalez, M. P.; Diaz, H. G. A topological sub-structural approach for predicting human intestinal absorption of drugs. *Eur. J. Med. Chem*. **2004**, *39*, 905-916.

63. Wegner, J. K.; Frohlich, H.; Zell, A. Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 931-939.

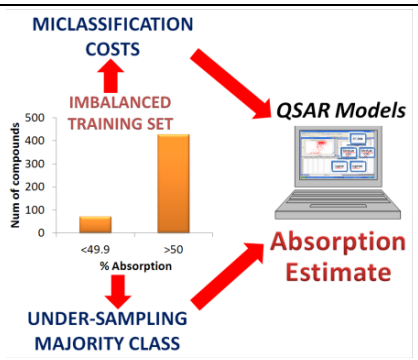
64. Yen, T. E.; Agatonovic-Kustrin, S.; Evans, A. M.; Nation, R. L.; Ryand, J. Prediction of drug absorption based on immobilized artificial membrane (IAM) chromatography separation and calculated molecular descriptors. *J. Pharm. Biomed. Anal*. **2005**, *38*, 472-478.

65. Van Hulse, J.; Khoshgoftaar, T. Knowledge discovery from imbalanced and noisy data. *Data Knowl. Eng*. **2009**, *68*, 1513-1542.

66. Zhang, Q. Y.; Hughes-Oliver, J. M.; Ng, R. T. A Model-Based Ensembling Approach for Developing QSARs. *J. Chem. Inf. Model*. **2009**, *49*, 1857-1865.

67. Li, Q. L.; Wang, Y. L.; Bryant, S. H. A novel method for mining highly imbalanced high-throughput screening data in PubChem. Bioinformatics. **2009**, *25*, 3310-3316.

1 **For Table of Contents Use Only**

	Coping with unbalanced class datasets in oral absorption models Danielle Newby, Alex. A. Freitas, Taravat Ghafourian*	

2